Peer-reviewed version available at *Energies* **2018**, *11*, 380; doi:10.3390/en11020380

Article

Bayesian Energy Measurement and Verification Analysis

Herman Carstens [1*], Xiaohua Xia [1] and Sarma Yadavalli [2]

[1] Centre for New Energy Systems, University of Pretoria
[2] Department of Industrial and Systems Engineering, University of Pretoria
* Correspondence: hermancarstens@gmail.com

Abstract: Energy Measurement and Verification (M&V) aims to make inferences about the savings achieved in energy projects, given the data and other information at hand. Traditionally, a frequentist approach has been used to quantify these savings and their associated uncertainties. We demonstrate that the Bayesian paradigm is an intuitive, coherent, and powerful alternative framework within which M&V can be done. Its advantages and limitations are discussed, and two examples from the industry-standard International Performance Measurement and Verification Protocol (IPMVP) are solved using the framework. Bayesian analysis is shown to describe the problem more thoroughly and yield richer information and uncertainty quantification than the standard methods while not sacrificing model simplicity. We also show that Bayesian methods can be more robust to outliers. Bayesian alternatives to standard M&V methods are listed, and examples from literature are cited.

Keywords: statistics; uncertainty; regression; sampling; outlier; probabilistic

---

1. Introduction

This study argues for the adoption of the Bayesian paradigm in energy Measurement and Verification (M&V) analysis by M&V practitioners and researchers. As such, no new Bayesian methods will be developed. Instead, the advantages, limitations, and application of the Bayesian approach to M&V will be explored. Since the focus is on application, a full explanation of the underlying theory of the Bayesian paradigm will not be given. Readers are referred to Kruschke [1] for a basic introduction, or Gelman [2] for a more advanced explanation.

The argument made below is not that current methods are completely wrong or that the Bayesian paradigm is the only viable option, but that the field can benefit from a greater adoption of Bayesian thinking because of its ease of implementation and accuracy of analysis.

This paper is arranged as follows. After giving a background on current M&V analysis methods and the opportunities for improvement in Section 1.1, the Bayesian paradigm is introduced and its practical benefits and limitations are discussed in Section 2. Section 3 offers two well-known examples and their Bayesian solutions. We also discuss robustness and hierarchical modelling. Section 4 gives a reference list of Bayesian solutions to common M&V cases.

1.1. Background

M&V is the discipline in which the savings from energy efficiency, demand response, and demand-side management projects are quantified [3], based on measurements and energy models. A large proportion of such M&V studies quantify savings for building projects, both residential and commercial. The process usually involves taking measurements or sampling a population to create a baseline, after which an intervention is done. The results are also measured, and the savings are inferred as the difference between the actual post-intervention energy use, and what it would have been, had no intervention taken place. These savings are expressed in probabilistic terms following the ISO Guide to the Expression of Uncertainty in Measurement (GUM) [4]. M&V study results often

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

form the basis of payment decisions in energy performance contracts, and the risk-implications of such studies are therefore of interest to decision makers.

The Bayesian option will not affect the basic M&V methodologies such as retrofit isolation or whole facility measurement, but only the way the data are analysed once one of these methods has been decided upon.

M&V guidelines such as the International Performance Measurement and Verification Protocol (IPMVP) [3], the American Society of Heating, Refrigeration, and Air Conditioning Engineers (ASHRAE)'s Guideline 14 on Measurement of Energy, Demand, and Water Savings [5], or the United States Department of Energy's Uniform Methods Project (UMP) [6], as well as most practitioners, use frequentist (or classical) statistics for analysis. Because of its popularity in the twentieth century, most practitioners are unaware that this is only one statistical paradigm and that its assumptions can be limiting. The term 'frequentist' derives from the method equating probability with long-run frequency. For coin flips or samples from a production line, this assumption may be valid. However, for many events, equating probability with frequency seems strained, because a large, hypothetical long-run population needs to be imagined for the probability-as-frequency-view to hold. Kruschke [1] gives an example where a coin is flipped twenty times and seven heads are observed. The question is then, what is the probability of the coin being fair? The frequentist answer will depend on the imagined population from which the data were obtained. This population could be "stopping after 20 flips", but it could also be "stopping after seven heads" or "stopping after two minutes of flipping" or "to compare it to another coin which was flipped twenty times". In each case the probability that it is a fair coin changes, even though the data did not – termed *incoherence* [7]. In fact, the probabilities are dependent on the analyst's *intention*. By changing his intention, he can alter the probabilities. This problem becomes even more severe in real-world energy savings inference problems with many more factors. The hypothetical larger population from which the energy use at a specific time on a specific day for a specific facility was sampled, is difficult to imagine. That is not to say that a frequentist statistical analysis cannot be done, or be useful. However, it often does not answer the question that the analyst is asking; an "error of the third kind". Analysts have become used to these 'statistical' answers (e.g. "not able to reject the null hypothesis"), and have accepted such confusion as part of statistics. For example, consider two mainstays of frequentist M&V: confidence intervals (CIs) and *p*-values. CIs are widely used in M&V to quantify uncertainty. According to Neyman, who devised these intervals, they do not convey a degree of belief, or confidence, as is often thought. They are a product of a process that produces an interval which contains the true value a given percentage of the time [8]. This may seem like practically the same thing, but consider Montgomery and Runger's notch-impact test example in their textbook *Applied Statistics and Probability for Engineers* [9], under "Interpreting a Confidence Interval". They explain CIs as follows (bold and italic emphases are theirs):

> How does one interpret a confidence interval? In the impact energy estimation problem in [the notch impact test] Example 8-1 the 95% CI is $63 \leq \mu \leq 65.08$J , so it is tempting to conclude that μ is within this interval with probability 0.95. However, with a little reflection, it's easy to see that this cannot be correct; the true value of μ is unknown and the statement $63 \leq \mu \leq 65.08$ is either correct (true with probability 1) or incorrect (false with probability 1). The correct interpretation lies in the realization that a CI is a *random interval* because the probability statement defining the end-points of the interval L and U [lower and upper] are random variables. Consequently, the correct interpretation of a ... CI depends on the relative frequency view of probability. Specifically, if an infinite number of random samples are collected and a [95%] confidence interval for μ is computed for each sample, [95%] of these intervals will contain the true value of μ.
>
> ...
>
> Now in practice, we obtain only one random sample and calculate one confidence interval. Since this interval either will or will not contain the true value of μ, it is not reasonable to attach a probability level to this specific event. An appropriate statement

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

> is the observed interval $[l, u]$ brackets the true value of μ with **confidence** [95%]. This statement has a frequency interpretation; that is, we don't know if the statement is true for this specific example, but the *method* used to obtain the interval $[l, u]$ yields correct statements [95%] of the time.

Consider now the p-value. Because of the confusion surrounding this statistic, the American Statistical Association issued a statement regarding its use [10], in which they say:

- P-values do not measure the probability that the studied hypothesis is true or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Such statements by professional statisticians leave most M&V practitioners confused, and rightly so. It is not that these methods are invalid, but that they have been co-opted to answer different *kinds* of questions to what they actually answer. The reason for their popularity in the 20th century has more to do with their computational ease, compared to the more formal and mathematical Bayesian methods, than with their appropriateness. The Bayesian conditional-probability paradigm is actually much older than the frequentist one but used to be impractical for computational reasons. However, with the rise in computing power and new numeric methods for solving Bayesian models, this is no longer a consideration.

2. The Bayesian Paradigm

Instead of approaching uncertainty in terms of long-run frequency, the Bayesian paradigm views uncertainty as a state of knowledge or a degree of belief; the sense most often meant by people when thinking about uncertainty. These uncertainties are calculated using conditional-probability logic and calculus, proceeding from first principles. For example, consider two conditions M and S. Let Pr denote a probability and | "conditional on" or "given". Bayes' theorem states that

$$\Pr(S|M) = \frac{\Pr(M|S)\Pr(S)}{\Pr(M)}. \tag{1}$$

Now, as stated previously, M&V is about verifying the savings achieved, based on some measurements and an energy model, and quantifying the uncertainty in this figure. If we let S be the savings, and M the measurements, Bayes' theorem as stated above answers that question exactly: it supplies a probability of the savings given the measurements; $\Pr(S|M)$. Bayes' theorem is, therefore, the natural expression of the M&V aim:

$$\text{Verification}|\text{Measurement} \equiv \Pr(S|M).$$

Whereas the frequentist paradigm views the data as random realisations of a process with fixed parameters, the Bayesian paradigm views the data (measurements) as known, and the underlying parameters as uncertain (thereby ensuring coherence [7]). This seems like a trivial distinction at first, but is significant: the frequentist only solves for $\Pr(M|S)$: the probability of observing that data, given the underlying savings value. However, that is not the question M&V seeks to answer. In the frequentist paradigm the analyst does not invert this as Bayes' theorem does to find the probability

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

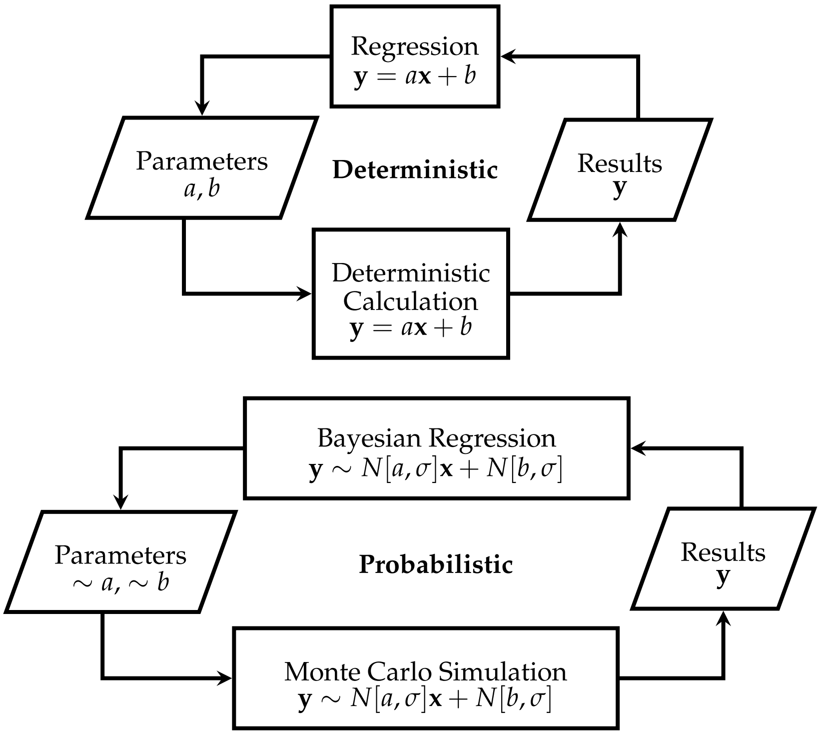

Figure 1. Deterministic and probabilistic calculation, simulation, and inverse modelling. The notation $\sim N[\cdot]$ denotes a normal distribution as a convenient substitute for any distribution.

distribution on the savings, given the data. That is the wrong question which is being answered, as alluded to above[1].

It is this inversion process that has often been intractable until the advent of Markov Chain Monte Carlo (MCMC) techniques and increased computing power. MCMC software has allowed users to specify a model (similar to a linear regression model, for example), supply the observations (measurements), and infer the values on the model parameters probabilistically. This is called probabilistic programming. Probabilistic programming is very powerful because instead of working with point estimates on all unknown parameters, one describes the system in terms of distributions. It is well known that doing calculations with point estimates of variables rather than with distributions on variables can be dangerous [11]. When doing forward-calculations as illustrated in Figure 1, it is therefore desirable to use distributions on unknown variables and then apply a Monte Carlo simulation or Mellin Transform Moment Calculation method [12,13] to obtain a probability distribution on the result. MCMC allows one to do the inverse: inferring parameter distributions from given data and a model. Therefore MCMC is to regression what Monte Carlo simulation is to deterministic computation. The adoption of the Bayesian paradigm therefore allows the analyst to move from deterministic to probabilistic M&V, as shown in Figure 1.

For the inversion described above to work, the $\Pr(S)$ term, called the prior, needs to be specified. Although the prior can be used to incorporate information into the model which is not available through the data alone, it is, in essence, merely a mathematical device allowing inversion. The prior is often specified as "non-informative" – a flat probability distribution over the region of interest, allowing the data to "speak for itself" through the likelihood term. This will be discussed in more

---

1 On a technical point, to be fair, we note that for non-informative priors, the likelihood may be equivalent to the posterior. This is not guaranteed, however. When it is the case, the frequentist likelihood may borrow from Bayesian theory and be interpreted as a probability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

detail below. The other term, $\Pr(M)$, need not be specified in numeric MCMC models – it is merely a normalising factor ensuring that the right-hand side of the equation can integrate to unity, making it a proper probability density function. The left-hand side of the equation is called the posterior distribution, and is proportional, therefore, to the product of the prior and the likelihood.

Advanced Bayesian models may be nuanced, but the fundamental mechanics as described above stay the same for all Bayesian analyses: specify priors, describe the likelihood, and solve to find the posterior on the parameters of interest.

2.1. Practical Benefits

Besides the theoretical attractiveness discussed above, the Bayesian paradigm also offers many practical benefits for energy M&V:

1. Because Bayesian models are probabilistic, uncertainty is automatically and exactly quantified.
2. Uncertainty calculations in the Bayesian approach can be much less conservative than standard approaches. Shonder and Im [14] show a 40% reduction in uncertainty in one case. Since project payment is often dependent on savings uncertainties being within certain bounds, using the Bayesian approach can increase project feasibility.
3. By making the priors and energy model explicit, the Bayesian approach ensures greater transparency – one of the five key principles of M&V [3].
4. The Bayesian approach is widely used and is rapidly gaining popularity in other scientific fields. Lira [15] relates that even the GUM (adopted by many societies of physics, chemistry, electrotechnics, etc.) is being rewritten to be more consistent with this approach. Since M&V reports uncertainty according to the GUM, Bayesian calculations would be useful.
5. Bayesian models are more universal and flexible than standard methods. Bayesian modelling can be highly sophisticated, but the core of probabilistic thinking is consistent throughout. This is different to frequentist statistics where knowledge of one or even many tests will not necessarily aid the analyst in understanding a new metric, or approach to a problem not seen before. Many frequentist tests are ad-hoc and apply only to specific situations. For example, t-tests have little to do with regression in frequentism, but in Bayesian thinking they are expressions of the same idea.
6. Being modular, Bayesian modelling is more flexible. Ordinary least squares (OLS) linear regression assumes residuals are normally distributed and that the variance is constant for all points. In a probabilistic Bayesian model the parameters can be distributed according to any distribution, but the posterior for each will be determined by the data (if the prior is appropriately chosen). Models are also modular, and can be designed to suit the problem. For example, it is no different to create terms for serial correlation, or heteroscedasticity (non-constant variance) than it is to specify an ordinary linear model. This also allows for easy specification of non-routine adjustments, the handling of missing values, and the incorporation of unmeasured yet important quantities such as measurement error; often problematic for energy models. For the retrofit isolation with key parameter measurement approach, the unmeasured parameters (the estimates) can also be incorporated in this way.
7. Bayesian models can account for model-selection uncertainty. There are often multiple reasonable energy models which could describe a specific case. For example time and dry-bulb temperature; occupancy and dry-bulb temperature; temperature, humidity, and occupancy, etc. The analyst usually chooses one model, discards the rest, and reports the uncertainty produced in that specific model. However, this uncertainty does not account for model selection. In other words, there is an uncertainty associated with choosing that specific model above another reasonable one. Bayesian model averaging allows many models to be specified simultaneously, and averages their results by automatically weighting each model's influence on the final result by that model's explanatory power. This gives a far more realistic uncertainty value [2].
8. Because uncertainty is automatically quantified, CIs can be interpreted in the way most people understand them: degrees of belief about the value of the parameter.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

9. The Bayesian approach is well-suited to "small data" problems. This seems like a minor point in developed countries where questions surrounding big data seem more pressing. However, big (energy) data is a decidedly "first-world problem". In developing countries a lack of meters makes M&V expensive, and it is useful to have a method that is consistent on smaller data sets as well.
10. The Bayesian approach allows for the incorporation of prior information where appropriate. The danger in this will be discussed in Section 2.2. However, in cases where it is warranted, known values or ranges for certain coefficients can be specified in the prior. This has been done successfully for energy projects [16–19]. Prior information is also useful in longitudinal studies, where measurements or samples from previous years can be taken into account [20,21].
11. When the savings need to be calculated for "normalised conditions", for example a 'typical meteorological year', rather than the conditions during the post-retrofit monitoring period, it is not possible to quantify uncertainty using current methods. However, Shonder and Im [14] have shown that it can be naturally and easily quantified using the Bayesian approach.
12. Bayesian approaches allow real-time or online updating of estimates [20–22]. For other machine learning techniques, the data need to be split into testing and training sets, the model trained on the training set, and then used to predict the testing set period. As new data becomes available, the model needs to be retrained in many cases[2], making it computationally expensive to keep a model updated. In a Bayesian paradigm, previous data can be summarised by the prior so that the model need not be retrained.

2.2. Limitations

The Bayesian approach also has limitations that M&V practitioners and policy makers should bear in mind.

1. Modelling is non-generic. In point 5 above it was stated that the Bayesian approach is more universal. This is true in the sense that the same basic approach is used for many different kinds of problems. However, the inherent modularity of the method as described in point 6 means that for most cases there is not a one-size-fits-all generic template in Bayesian modelling, the way there usually is in frequentist modelling. This requires a bit more thinking from the analyst. However, we believe this to be an advantage: frequentist approaches make it easier to think less, but as a consequence, also to build poor models, which has led to the current replication crisis seen in research [23] and a general mistrust of statistical results [24]. High quality models require some thought and care, in any paradigm.
2. As with any method, it is not immune to abuse. The most popular criticism is that by having a prior distribution on the savings, the posterior may be biased in a way not warranted by the data, making the result subjective. This is certainly possible. However, having a prior in an M&V analysis is actually an advantage.
   (a) As stated above, it allows for greater modelling transparency. The Bayesian form forces the analyst to be explicit about his or her modelling assumptions, and to defend them. Such assumptions cannot be imported by (accidentally or purposefully) choosing one test over another, as in the frequentist case.
   (b) It is sometimes necessary to include priors to *avoid* bias. Ioannidis [25] and Button [26] have shown that many medical studies contain false conclusions due to biased results. The bias that was introduced was to consider positive and negative outcomes from a clinical trial equally likely. However, the prior odds of an experimental treatment working is much lower than the odds of that treatment not working. Ignoring these prior odds leads to a high false-positive rate, since many of the positive results are actually false – due to noise. In

[2] Artificial Neural Networks (ANNs) are an exception.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

M&V the situation is reversed: the prior odds of energy projects saving energy is generally high. Having a neutral prior would therefore bias a result towards conservatism; one of the key principles of M&V [3]. Nevertheless, Button's study is an excellent illustration of why priors are an important part of probability calculus.

(c) Because the assumptions and distributions used are clearly stated, it precludes hedging the M&V result with phrases such as "however, from previous studies/experience we know that this is a conservative figure...".

(d) The thorough analyst will test the effect of different priors on the posterior, demonstrating the bias introduced through his modelling assumptions, and justifying its use.

3. Bayesian methods can be computationally expensive for large datasets and complex models. It is true that numerical solvers are becoming more efficient and computational power is increasing. However, in comparison with matrix inversion techniques used for linear regression, for example, Bayesian methods are much slower and may be inappropriate for real-time applications [27].
4. The forecasting accuracy of other machine learning methods is higher in some cases. Some machine learning (ML) techniques such as ANNs or Boosted trees are more accurate than regression-type approaches in some cases [28,29], although regression-based approaches such as time-of-week-and-temperature [30] still perform very well [29,31] and may be preferred for simplicity. It also depends on the problem: it is not possible to know beforehand which model will work the best [32]. ML algorithms also still only produce point estimates. Therefore they cannot be compared to the full probabilistic approach which provides much richer information and is not just a forecasting technique, but a full inference paradigm. However, accuracy is still important. Bayesian ANN packages have been developed recently [33], and show great promise for combining the best of both approaches.
5. The parametric from of the model needs to be specified. Parametric Bayesian models as described in most of this study is that they can only be correct in so far as their functional form describes the underlying physical process. Model misspecification is a real possibility. This is different to the machine learning methods described in the previous paragraph, which do not rely on a functional form being specified. Non-parametric models have their own benefits and limitations: for cases where the underlying physical process is well-understood, a parametric model can be more accurate. Non-parametric methods also have their own set of assumptions that need to be satisfied. Nevertheless, Gaussian Processes (GPs) are non-parametric Bayesian methods which do not suffer from the model misspecification risk, and have been applied successfully to energy M&V problems [16,34] which are often complex and defy functional descriptions. Gaussian Mixture Models have also been applied [35].

3. Bayesian M&V Examples

To demystify the Bayesian approach, two basic M&V calculations will be demonstrated. The reader will notice the recurring theme of expressing all variables as (conditional) probability distributions.

3.1. Sampling Estimation

Consider the following example from the IPMVP 2012 [3, Appendix B-1]. Twelve readings are taken by a meter. These are reported as monthly readings, but are assumed to be uncorrelated with any independent variables or other readings, and are therefore construed to be random samples. The values are

$$\mathbf{D} = [950, 1090, 850, 920, 1120, 820, 760, 1210, 1040, 930, 1110, 1200]. \quad (2)$$

The units are not reported and the results below are therefore left dimensionless, although kWh would be a reasonable assumption. These data were carefully chosen, and have a mean $\mu = 1\,000$, sample standard deviation $s_s = 150$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

3.1.1. IPMVP solution

The standard error is $SE = 43$. The confidence interval on the mean is calculated as

$$CI = \mu \pm t \times SE \tag{3}$$

Since $t_{90\%,11} = 1.80$, the 90% confidence interval on the mean was calculated as $1000 \pm 1.80 \times 43 = (933,\ 1\,077)$, or a 7.7% precision. Metering uncertainty is not considered in this calculation.

3.1.2. Bayesian solution

The Bayesian estimate of the mean is calculated as follows. First, prior distributions on the data need to be specified. Vague priors will be used:

$$\Pr(\mu) \sim Uniform[0, 2000], \tag{4}$$

$$\Pr(\sigma) \sim Uniform[0, 1000]. \tag{5}$$

A t distribution will be used for the likelihood below, and the degrees of freedom parameter (ν) of this distribution will, therefore, need to be specified. One could fix ν for the t-distribution at 12, since there are twelve data points and traditionally ν has been taken to signify this number. However, if outliers are present or if the data has more or less dispersion than the standard t-distribution with as many data points, this would not be realistic. It is therefore warranted to indicate the uncertainty in the data by specifying a prior distribution on ν also: a hyperprior. Kruschke [36] recommends an exponential distribution with the mean equal to the number of data points. This allows equal probability of ν being higher or lower than the default value:

$$\Pr(\nu) \sim Exponential[1/12]. \tag{6}$$

If the vector of the parameters is $\boldsymbol{\theta} = (\mu, \sigma, \nu)$, then the likelihood can be written as:

$$\Pr(\mathbf{D}|\boldsymbol{\theta}) \sim StudentT\left[\Pr(\mu),\ \Pr(\sigma),\ \Pr(\nu)\right]. \tag{7}$$

Note that the t distribution is not used because of the t-test, but because its heavier tails are more accommodating of outliers. Any distribution could have been specified if there was good reason to do so. The posterior on μ is plotted in Figure 2. It was simulated in PyMC3 using the ADVI algorithm with 100 000 draws, which is stable and converges on the posterior distribution in 10.76 seconds on a middle-range laptop computer.Although this notation may seem intimidating to practitioners who are not used to it, writing this in the probabilistic Python programming package PyMC3 [37] demonstrates the intuitive nature of such a model:

```python
import pymc3 as pm
with pm.Model() as bayesian_sampling_model:
   # Hyperpriors and priors:
   mean = pm.Uniform('mean', 0, 2000)
   std = pm.Uniform('std', 0, 1000)
   nu = pm.Exponential('nu', 1/len(data))
   #  Likelihood
   likelihood = pm.StudentT('likelihood', mu=mean, sd=std, nu=nu, observed=data)
   # ADVI calculation
   trace = pm.variational.sample_vp(vparams=pm.variational.advi(n=100000))
```

It is important to note that no probability statements about the values inside the frequentist interval can be made, nor can one fit a distribution to the interval. The distribution indicated is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1
Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

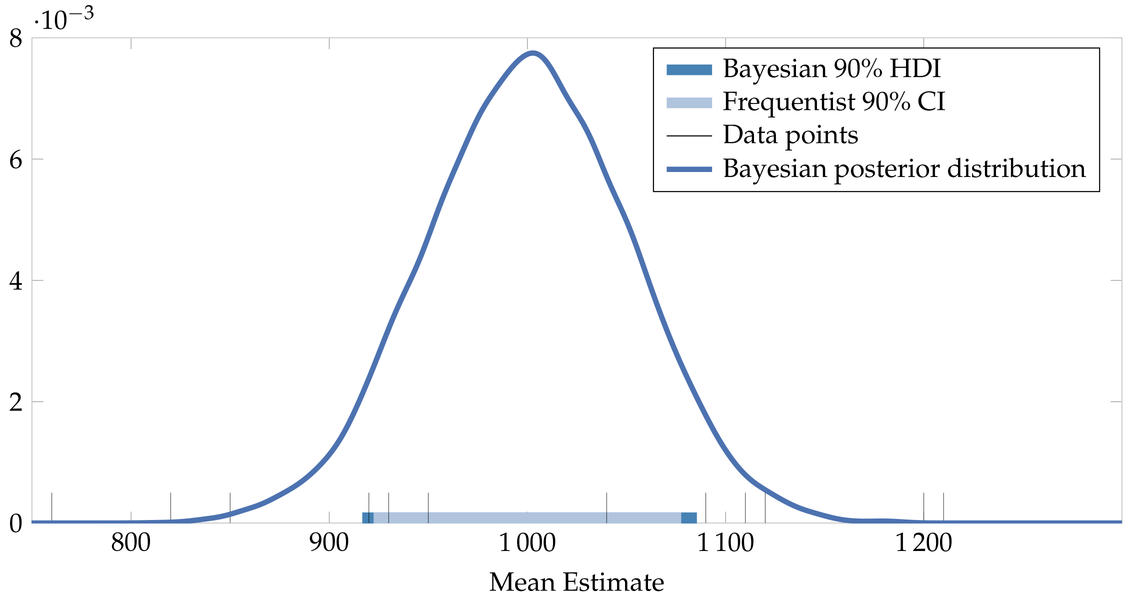

Figure 2. Illustration of Bayesian posterior density $\Pr(\mu|\mathbf{D})$, 90% HDI, and frequentist 90% CI.

strictly a Bayesian one. The Bayesian (highest density) interval is slightly wider than the frequentist confidence interval, at a precision of 8.5%. If ν were fixed at 12, (indicating that we are certain that the data does indeed reflect a t distribution with 12 degrees of freedom exactly), Bayesian and frequentist intervals correspond exactly. However, the Bayesian alternative allows for a more realistic value. With comparisons between two groups (two-sample t-tests), the effect of uncertainty in the priors becomes even more pronounced [36].

The posterior distribution can now be used to answer many interesting questions. For instance, what is the probability, given the data at hand, that the true mean is below 900? Or, is it safe to assume that the standard value of 950 is reflected by this sample, or should the null hypothesis be rejected? (If previous data to this effect is available, it could be included in the prior, maybe using the equivalent prior sample size method [38]). The frequentist may say that there is not enough evidence to reject the null, but cannot accept it either. In the Bayesian paradigm, 950 falls comfortably within the 90% confidence range, and can therefore be accepted at that level. As a further question, if this is an energy performance contracting project, and we assume that the data points are different facilities rather than different months, would it be worthwhile taking a larger sample to increase profits, if we believe that the true mean is at 1 100? (On which see Lindley [39], Bernardo [40] and Goldberg [41]).

It is therefore evident that the Bayesian result yields richer and more useful information using intuitive mathematics.

3.2. Regression

In M&V, one often uses the baseline data ($\mathbf{D}_b$) to infer the baseline (pre-retrofit) model parameters $\boldsymbol{\theta}$ through an inverse method:

$$\boldsymbol{\theta} = f^{-1}(\mathbf{D}_b, \tau), \tag{8}$$

Where $f(\cdot)$ is a function relating the independent variables (energy governing factors) to the energy use of the facility, and τ is time. The model parameters describe the sensitivity of the energy model to the independent variables such as occupancy, outside air temperature, or production volume.

As an aside, this section will discuss an elementary parametric energy model using Bayesian regression, similar to standard linear regression. In practice, a two-parameter linear regression model seldom captures the different states of a facility's energy use, for example, heating at low temperatures,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

a comfortable range, and cooling at high temperatures. Piecewise linear regression techniques are often used [42–46], and they tend to work reasonably well if their assumptions are satisfied, but they are not stable in all cases, are approximate, and the assumptions are often restrictive. Shonder and Im [14] provide a Bayesian alternative. A non-parametric model using a Gaussian Process could also be used, and since one does not need to specify a parametric model, it allows very flexible models to be fit while still quantifying uncertainty. This is especially useful for models where energy use is a nonlinear function of the energy governing factors. However, to keep the example simple and focussed, only a simple parametric model will be considered below.

3.2.1. Example

Suppose one has a simple regression model where the energy use of a building **E** is correlated with the outside air temperature through the number of Cooling Degree Days (CDD). One cooling degree day is defined as an instance where the average daily temperature is one degree above the thermostat set point, and the building therefore requires one degree of cooling. Let the intercept coefficient be θ_0, the slope coefficient θ_1, and the Gaussian error term ϵ. One could then write

$$\mathbf{E} = \theta_0 + \theta_1 \mathbf{CDD} + \epsilon. \tag{9}$$

In standard linear regression, one would write $\hat{\boldsymbol{\theta}}$ as the vector of two coefficients and do some linear algebra to obtain their estimates. There would be a standard error on each, which would indicate their uncertainties, and if the assumptions of linear regression, such as normality of residuals, independence of data, homoscedasticity, etc. hold, then it would be accurate. In Bayesian regression, one would describe the distributions on the parameters

$$\Pr(\boldsymbol{\theta}|\mathbf{D}) \propto \Pr(\mathbf{D}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) \sim N[\hat{\boldsymbol{\theta}}, \boldsymbol{\sigma}] \tag{10}$$

where $\boldsymbol{\sigma}$ is the vector of the standard deviations on the estimates. Generating random pairs of values from the posterior, at a given value of CDD, according to the appropriate distributions, will yield the posterior predictive distribution. This is the distribution of energy use at a given temperature, or over the range of temperatures. Overlaying such realisations onto the actual data is called the posterior predictive check.

Now consider a concrete example. The IPMVP 2012 [3, Appendix B-6] contains a simple regression example of creating a baseline of a building's cooling load. The twelve data points themselves were not given, but a very similar data set yielding almost identical regression characteristics has been engineered and is shown in Table 1.

Table 1. Cooling Degree Day Data for IPMVP Example B-6. Note that these data were reverse-engineered to yield the same regression results as reported in the IPMVP. The original data were not reported in the IPMVP.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
CDD	312	292	222	112	92	22	12	32	157	207	182	302
Energy Use	7823	7585	7486	6646	6185	5933	5381	5917	7158	7064	7231	8250

A linear regression model was fit to the data, and yielded the result shown in Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

Table 2. Linear regression fit characteristics for data in Table 1. The coefficient of determination is $R^2 = 0.93$, which is identical to the IPMVP case. These results may be compared to Bayesian summary statistics in Table 3.

Parameter	Value	Standard Error	95% Interval
Slope coefficient	7.75	0.67	[6.26, 9.23]
Intercept coefficient	5634	129	[5347, 5921]

3.2.2. IPMVP Solution

The IPMVP then proceeds to calculate the uncertainty in the annual energy figure by multiplying the standard error on the estimate (the average standard error) by $t_{95\%}$ and the average consumption in the average month, and assumes that this value is constant for all months. As discussed in this study, this approach is problematic, and can at best be seen as approximate. Since it is treated in some detail in the IPMVP, the analysis will not be repeated here.

3.2.3. Bayesian Solution

The key to the Bayesian method is to approach the problem probabilistically, and therefore view all parameters in (9) as probability distributions, and specify them as such. In this regression model there are three parameters of interest: the intercept (θ_0), slope (θ_1), and the response (**E**). This response is the likelihood function, familiar to most readers as the frequentist approach. Each of these distributions need to be specified in the Bayesian model. First, consider the priors on the slope and intercept. These can be vague[3]:

$$\Pr(\theta_0) \sim Uniform[0, 10000], \tag{11}$$

and

$$\Pr(\theta_1) \sim Uniform[0, 20]. \tag{12}$$

Now consider the likelihood. In frequentist statistics one needs to assume that **E** in (9) is normally distributed. In the Bayesian paradigm one may do so, but it is not necessary. A Student's *t*-distribution is often used instead of a Normal distribution in other statistical calculations (e.g. *t*-tests) due to its additional ("degrees of freedom") parameter which accommodates the variance arising from small sample sizes more successfully. As in Section 3.1.2, an exponential distribution on the degrees of freedom (ν_p) is specified. It has also been found that specifying a Half-Cauchy distribution on the standard deviation (σ_p) works well [48]. Therefore the hyperpriors are specified as

$$\Pr(\nu_p) \sim Exponential[12^{-1}] \tag{13}$$

and

$$\Pr(\sigma_p) \sim HalfCauchy[1]. \tag{14}$$

The mean of the likelihood is the final hyperparameter that needs to be specified. This is simply (9), written with the priors:

$$\mu_p = \Pr(\theta_0) + \Pr(\theta_1)\mathbf{CDD}. \tag{15}$$

The full likelihood can thus be written as

$$\Pr(\mathbf{CDD}|\mathbf{E}) \sim StudentT\left(\mu = \mu_p, \nu = \Pr(\nu_p), \sigma = \Pr(\sigma_p)\right). \tag{16}$$

The PyMC3 code is shown below:

[3] Strictly speaking one can specify more scale-invariant priors than simply using normal or uniform distributions in $Pr(\boldsymbol{\theta})$ [47]. However, in practice we have not seen this done.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380


```
import pymc3 as pm
with pm.Model() as bayesian_regression_model:
    # Hyperpriors and priors:
    nu = pm.Exponential('nu', lam=1/len(CDD))
    sigma = pm.HalfCauchy('sigma', beta=1)
    slope = pm.Uniform('slope', lower=0, upper=20)
    intercept = pm.Uniform('intercept', lower=0, upper=10000)
    # Energy model:
    regression_eq = intercept + slope*CDD
    # Likelihood:
    y = pm.StudentT('y', mu=regression_eq, nu=nu, sd=sigma, observed=E)
    # MCMC calculation:
    trace = pm.sample(draws=10000, step=pm.NUTS(), njobs=4)
```

The last line of the code above invokes the MCMC sampler algorithm to solve the model. In this case the No U-Turn Sampler (NUTS) [49] was used, running four traces of 10 000 samples each, simultaneously on a four-core laptop computer, in 3.5 minutes. Fewer samples could also have been used.

A discussion of the inner workings and tests for adequate convergence of the MCMC is beyond the scope of the study and has been done in detail elsewhere in literature [2]. The key idea for M&V practitioners is that the MCMC, like MC simulation, must converge, and must have done enough iterations after convergence to approximate the posterior distribution numerically. For most simple models such as the ones used in most M&V applications, a few thousand iterations are usually adequate for inference. Two popular checks for posterior validity are the Gelman-Rubin statistic $\hat{R}$ [50,51] and the effective sample size (ESS). The Gelman-Rubin statistic compares the four chains specified in the program above, started at random places, to see if they all converged on the same posterior values. If they did, their ratios should be close to unity. This is easily done in PyMC3 with the `pm.gelman_rubin(trace)` command, which indicates $\hat{R}$ equal to one to beyond the third decimal place. However, even if the MCMC has converged, it does not mean that the chain is long enough to approximate the posterior distribution adequately because the MCMC mechanism produces a serially correlated (autocorrelated) chain. It is therefore necessary to calculate the *effective* sample size: the sample size taking autocorrelation into account. In PyMC3, one can invoke the `pm.effective_n(trace)` command, which shows that the ESSs for the parameters of interest are well over 1 000 each. As a first-order approximation, we can therefore be satisfied that the MCMC has yielded satisfactory estimates.

The MCMC results can be inspected in various ways. The posteriors on the parameters of interest are shown in Figure 3. If a normal distribution is specified on the likelihood in (16) rather than the Student's t, the posterior means are identical to the linear regression point estimates – an expected result, since OLS regression is a special case of the more general Bayesian approach. Using a t-distributed likelihood yields slightly different, but practically equivalent, results. The mean or mode of a given posterior is not of as much interest as the full distribution, since this full distribution will be used for any subsequent calculation. However, the mean of the posterior distributions are given in Table 3 for the curious reader.

Two brief notes on Bayesian intervals are necessary. As discussed in Section 1.1, the frequentist 'confidence' interval is a misnomer. To distinguish Bayesian from frequentist intervals, Bayesian intervals are often called 'credible' intervals, although they are much closer to what most people think of when referring to a frequentist confidence interval. The second note is that Bayesians often use Highest Density Intervals (HDIs), also known as highest posterior density intervals. These are related to the *area* under the probability density curve, rather than points on the x-axis. In frequentist statistics, we are used to equal-tailed confidence intervals since we compute them by taking the mean, and then adding or subtracting a fixed number - the standard error multiplied by the t-value,

Peer-reviewed version available at *Energies* **2018**, *11*, 380; doi:10.3390/en11020380

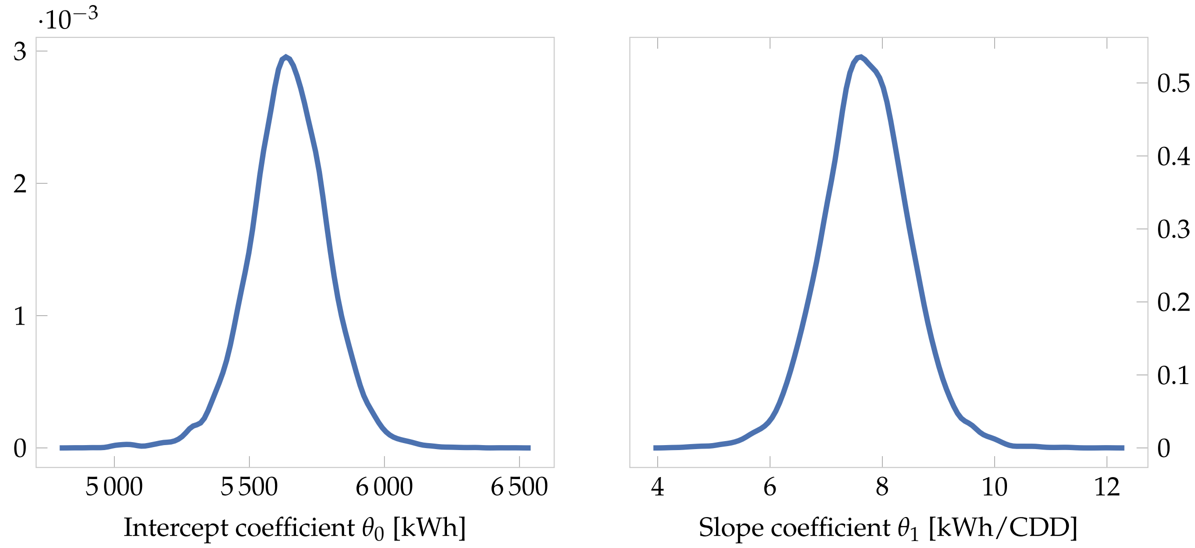

Figure 3. Posterior distributions on the parameters of interest. The summary statistics are given in Table 3.

for example. This works well for symmetrical distributions such as the Normal, as is assumed in many frequentist methods. However, real data distributions are often asymmetrical, and forcing an equal-tailed confidence interval onto an asymmetric distribution leads to including an unlikely range of values on the one side, while excluding more likely values on the other. An HDI solves this problem. It does not have equal tails, but has equally-likely upper and lower bounds.

Table 3. Summary statistics for Bayesian posterior distributions shown in Figure 3 when a Student's *t* distribution is used on the likelihood. Compare to linear regression results in Table 2.

Parameter	**Value**	**95% HDI**
Slope coefficient	7.69	[6.21, 9.24]
Intercept coefficient	5634	[5351, 5937]

The posterior distributions shown in Figure 3 are seldom of use in themselves and are more interesting when combined in a calculation to determine the uncertainties in the baseline as shown in Figure 4 or adjusted baseline. To do so the posterior predictive distribution needs to be calculated using the `pm.sample_ppc()` command, which resamples from the posterior distributions, much like the MC simulation forward-step of Figure 1.

The Bayesian model can also be used to calculate the *adjusted* baseline, or what the post-implementation period energy use would have been, had no intervention been made. The difference between this value and the actual energy use during the reporting period is the energy saved. For the example under consideration, the IPMVP assumes that an average month in the post-implementation period: one with 162 CDDs. It also assumes that the actual reporting period energy use is 4300 kWh, measured with negligible metering error. To calculate the savings distribution using the Bayesian method, one would do an MC simulation of

$$Savings \sim \boldsymbol{\theta}_0 + 162\boldsymbol{\theta}_1 - 4300 \tag{17}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are the distributions shown in Figure 3. Running this simulation with 10 000 samples yields the distribution shown in Figure 5. The 95% HDI is [2229, 2959], while the frequentist interval is

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

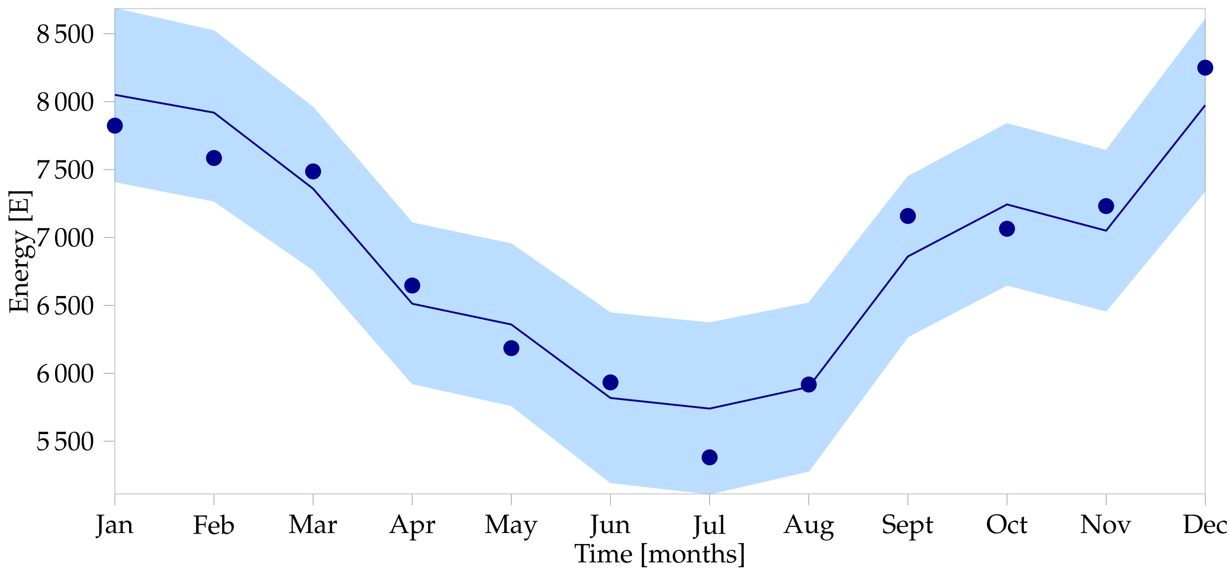

Figure 4. Measured data with overlaid Bayesian baseline model and its 95% Highest Density Interval.

[1810, 3430] for the same data – a much wider interval. Furthermore, the IPMVP then assumes averages and multiplies these figures to get annual savings and uncertainties. In the Bayesian paradigm, the HDIs can be different for every month (or time step) as shown in Figure 4, yielding more accurate overall savings uncertainty values.

3.2.4. Robustness to Outliers

As alluded to above, using the Student's t distribution rather than the normal distribution allows for Bayesian regression to be robust to outliers [52]. The heavier tails more easily accommodate an outlying data point by automatically altering the degrees-of-freedom hyperparameter to adapt to the non-normally distributed data. This feature does not give the the M&V practitioner free reign to ignore outliers. One should always seek to understand the reason for such an outlier; if the operating conditions of the facility were significantly different, it would be good modelling practice to neglect (or 'condone') the data point. However, it is not always possible to trace the reasons for all outliers, and inherently robust models are useful.

To demonstrate the robustness of such a Bayesian model, consider the regression case above. Suppose that for some reason the December cooling load was 3250 kWh and not 8250 kWh, indicated by the red point in the lower right hand corner of Figure 6. If OLS regression were used, and this point is not removed, it would skew the whole model. However, the t-distributed likelihood in the Bayesian model is robust to the outlier. The effect is demonstrated in Figure 6. Four lines are plotted: the solid lines are for the data set without the outlier. The dashed lines are for the data set with the outlier. In the Bayesian model the two regression lines are almost identical and close to the OLS regression line for the standard set. However, the OLS regression on the outlier set is dramatically biased and would underestimate the energy use for hot months due to the outlier.

3.2.5. Hierarchical Models

A further advantage in the Bayesian paradigm is the use of hierarchical, or multilevel models. This is a feature of the model structure rather than the Bayesian calculation itself (it also works for MLE) [1], but it is nevertheless useful in M&V. Suppose that multiple measures are installed at multiple sites so that the IPMVP Option C: Whole Building Retrofit is used for M&V. The UMP Chapter 8 [53]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at Energies 2018, 11, 380; doi:10.3390/en11020380

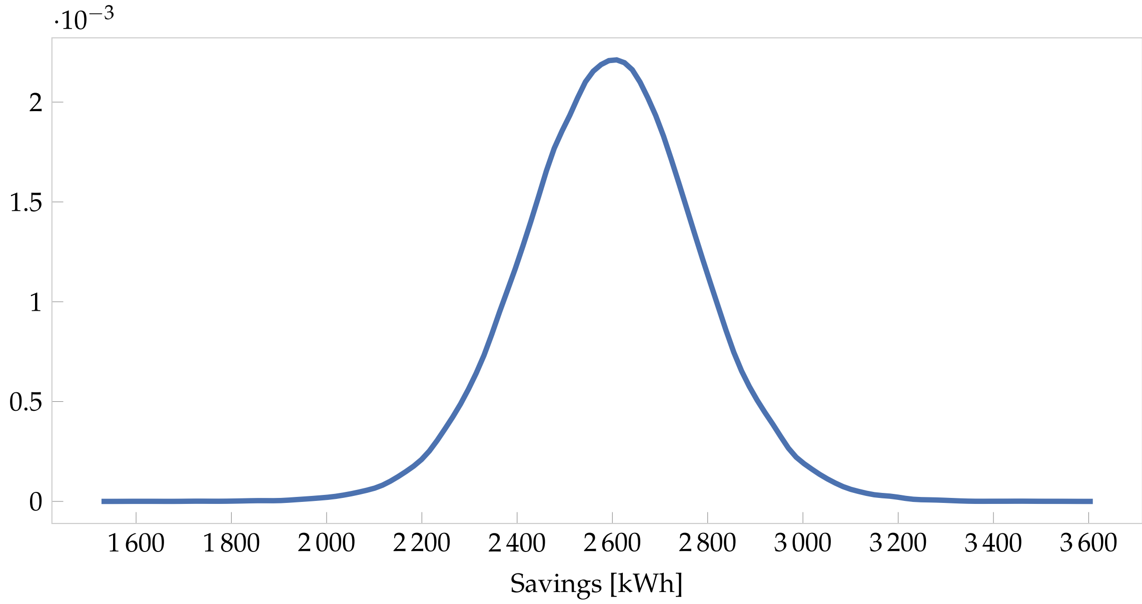

Figure 5. Distribution on the savings for a month with 162 CDDs.

reports that there are two ways to analyse such data. The two-stage approach involves first analysing each facility separately and then using these results for the overall analysis in stage two. The fixed effects approach analyses all buildings simultaneously but assumes that the effect sizes are constant across facilities, using an average effect for all buildings. Hierarchical modelling considers both the individual facility's energy saving and the overall effect simultaneously. It does this by assuming that the group effects are different realisations of an overarching distribution with a mean and variance, which is used as a prior. This can lead to 'shrinkage' because the group effects are mutually informative. For groups with little data, the overarching effect distribution plays a larger role, and for groups with more data, a smaller role. Also, the overall variance is reduced because the sources of inter-facility variance are isolated from that of inter-measure variance. The result for a hierarchical model is that the effect estimation for an individual facility is influenced by the overall estimate of the measured effect, as well as by the data for the facility. As another example, consider a program that retrofits air conditioning units in different provinces in South Africa. One could fix the savings effect across all facilities, but this will underestimate some and overestimate others. Alternatively one could analyse by facility, then by province, and then overall. The hierarchical model provides a better alternative in these cases, and comprises the bulk of many Bayesian data analysis texts [1,2]. Booth, Choudhary, and Spiegelhalter have provided an excellent example of using hierarchical Bayesian models in energy M&V [54].

4. Bayesian Alternatives for Standard M&V Analyses

At this point an M&V want to try the Bayesian method for an M&V problem, but where to start? In Table 4, some Bayesian alternatives to standard M&V analyses are given. The references cited are mostly from M&V studies, although some general statistical sources are also listed where applicable.

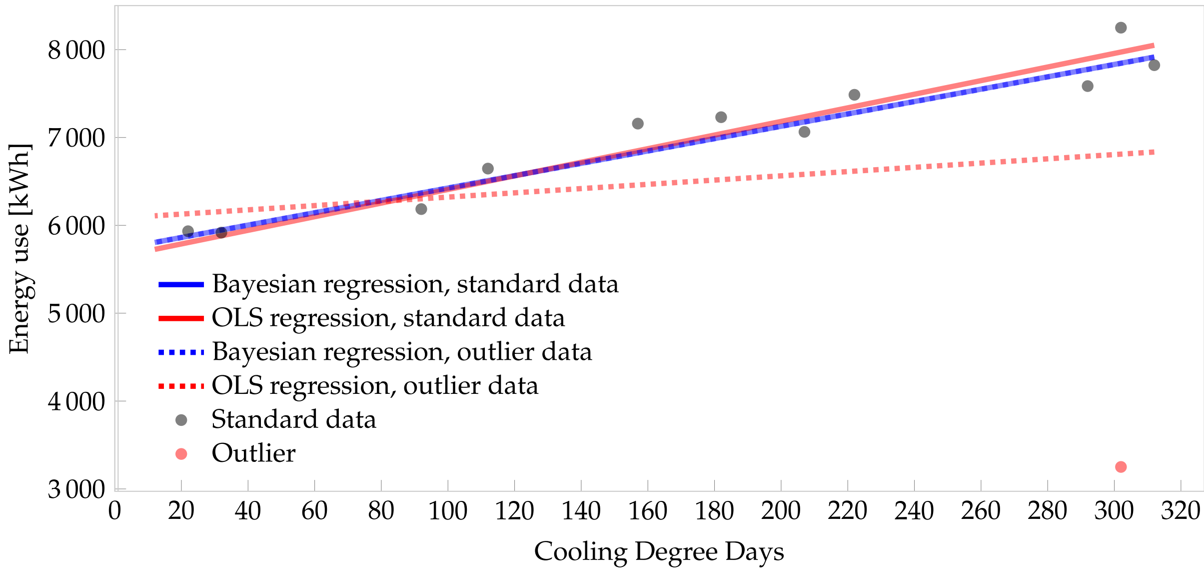

Figure 6. Demonstration of robustness of *t*-distributed Bayesian regression. Note that the two Bayesian regression lines (solid and dashed) coincide almost perfectly.

Table 4. Common M&V cases and their Bayesian alternatives

Problem type	Variant	Bayesian Alternative	Example Reference
Sampling	Single Sample		Section 3.1, [1]
	Randomised Control Trial	Bayesian Estimation	[36]
	ANOVA	Hierarchical modelling	[55]
Regression	Standard	Bayesian regression	Section 3.2, [22]
	With change points	Bayesian regression	[14]
	Pooled fixed effects	Hierarchical modelling	[54]
	Non-parametric	Gaussian Process	[34,56,57]
Longitudinal	Persistence	Dynamic Generalised Linear Model	[20]
Meter calibration		Simulation Extrapolation with Bayesian refinement	[58]

5. Conclusions

The Bayesian paradigm provides a coherent and intuitive approach to energy measurement and verification. It does so by defining the basic M&V question – the savings inference given measurements – using conditional probabilities. It also provides a simpler and more intuitive understanding of probability and uncertainty because it allows the analyst to answer real questions in a straightforward manner, unlike traditional statistics. Due to recent technological and mathematical advances being incorporated into software, analysts need not be expert statisticians to harness the power and flexibility of this method.

The probabilistic nature of Bayesian analysis allows for automatic and accurate uncertainty quantification in savings models. The richer nature of the Bayesian result is shown in a sampling and a regression problem, where it is found that the Bayesian method allows for more realistic modelling and a greater variety of questions that can be answered. Its flexibility is also demonstrated by constructing a robust regression model, which is much less sensitive to outliers that standard ordinary least squares regression traditionally used in M&V.

Acknowledgments: The authors would like to thank Dr Mark Rawlins for his input to the draft of this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at *Energies* **2018**, *11*, 380; doi:10.3390/en11020380

Author Contributions: Herman Carstens conceived of and wrote this paper. Xiaohua Xia and Sarma Yadavalli led the study and his doctoral work in which some of these ideas were developed.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
ASHRAE	American Society of Heating, Refrigeration, and Air Conditioning Engineers
CI	Confidence Interval
ESS	Effective Sample Size
GP	Gaussian Process
HDI	Highest Density Interval
IPMVP	International Performance Measurement and Verification Protocol
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
M&V	Measurement and Verification
OLS	Ordinary Least Squares
UMP	Uniform Methods Project

References

1. Kruschke, J. *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan*, 2 ed.; Academic Press, 2015.
2. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Vol. 2, Taylor & Francis, 2014.
3. Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol Vol. 1*, 2012.
4. Guide 98–3 (2008) Uncertainty of measurement Part 3: Guide to the expression of uncertainty in measurement (GUM: 1995). *ISO, Geneva* **2008**.
5. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc.. *Guideline 14-2014, Measurement of Energy, Demand, and Water Savings*, 2014.
6. National Renewable Energy Laboratory. *Uniform Methods Project*.
7. Robert, C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*; Springer Science & Business Media, 2007.
8. Neyman, J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **1937**, *236*, 333–380.
9. Montgomery, D.C.; Runger, G.C. *Applied Statistics and Probability for Engineers*; John Wiley & Sons, 2010.
10. Wasserstein, R.L.; Lazar, N.A. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* **2016**, *70*, 129–133.
11. Savage, S.L.; Markowitz, H.M. *The flaw of averages: Why we underestimate risk in the face of uncertainty*; John Wiley & Sons, 2009.
12. Kuang, Y.C.; Rajan, A.; Ooi, M.P.L.; Ong, T.C. Standard uncertainty evaluation of multivariate polynomial. *Measurement* **2014**, *58*, 483–494.
13. Rajan, A.; Ooi, M.P.L.; Kuang, Y.C.; Demidenko, S.N. Analytical Standard Uncertainty Evaluation Using Mellin Transform. *Access, IEEE* **2015**, *3*, 209–222.
14. Shonder, J.A.; Im, P. Bayesian Analysis of Savings from Retrofit Projects. *ASHRAE Transactions* **2012**, *118*, 367.
15. Lira, I. The GUM revision: the Bayesian view toward the expression of measurement uncertainty. *European Journal of Physics* **2016**, *37*, 025803.
16. Heo, Y. Bayesian Calibration of Building Energy Models for Energy Retrofit Decision-making under Uncertainty. PhD thesis, Georgia Institute of Technology, 2011.
17. Lee, B.D.; Sun, Y.; Augenbroe, G.; Paredis, C.J. Toward Better Prediction of Building Performance: a Workbench to Analyze Uncertainty in Building Simulation. BS2013: 13th Conference of the International Building Performance Simulation Association; , 2013.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at *Energies* **2018**, *11*, 380; doi:10.3390/en11020380

18. Booth, A.; Choudhary, R. Decision making under uncertainty in the retrofit analysis of the UK housing stock: Implications for the Green Deal. *Energy and Buildings* **2013**, *64*, 292–308.
19. Heo, Y.; Graziano, D.J.; Guzowski, L.; Muehleisen, R.T. Evaluation of calibration efficacy under different levels of uncertainty. *Journal of Building Performance Simulation* **2015**, *8*, 135–144.
20. Carstens, H.; Xia, X.; Yadavalli, S. Efficient Longitudinal Population Survival Survey Sampling for the Measurement and Verification of Building Retrofit Projects. *Energy and Buildings*, *150*, 163–176.
21. Carstens, H.; Xia, X.; Yadavalli, S. Efficient metering and surveying sampling designs in longitudinal Measurement and Verification for lighting retrofit. *Energy and Buildings*, *154*, 430–447.
22. Tehrani, N.H.; Khan, U.T.; Crawford, C. Baseline load forecasting using a Bayesian approach. 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2016, pp. 1–4.
23. McShane, B.B.; Gal, D.; Gelman, A.; Robert, C.; Tackett, J.L. Abandon statistical significance. *arXiv preprint arXiv:1709.07588* **2017**.
24. Leek, J.; Colquhoun, D.; McShane, B.B.; Gelman, A.; Nuijten, M.B.; Goodman, S.N. Comment: Five Ways to Fix Statistics. *Nature* **2017**, *551*, 557–559.
25. Ioannidis, J.P. Why most published research findings are false. *PLoS Med* **2005**, *2*, e124.
26. Button, K.S.; Ioannidis, J.P.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.; Munafò, M.R. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **2013**, *14*, 365–376.
27. Pavlak, G.S.; Florita, A.R.; Henze, G.P.; Rajagopalan, B. Comparison of Traditional and Bayesian Calibration Techniques for Gray-Box Modeling. *Journal of Architectural Engineering* **2013**, *20*, 04013011–2–16.
28. Yildiz, B.; Bilbao, J.; Dore, J.; Sproul, A. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy* **2017**.
29. Gallagher, C.V.; Bruton, K.; Leahy, K.; O'Sullivan, D.T. The Suitability of Machine Learning to Minimise Uncertainty in the Measurement and Verification of Energy Savings. *Energy and Buildings* **2017**.
30. Mathieu, J.L.; Price, P.N.; Sila, K.; Piette, M.A. Quantifying changes in building electricity use, with application to demand response. *IEEE Transactions on Smart Grid* **2009**, *41*, 374–381.
31. Granderson, J.; Touzani, S.; Custodio, C.; Sohn, M.D.; Jump, D.; Fernandes, S. Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings. *Applied Energy* **2016**, *173*, 296–308.
32. Wolpert, D.H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* **1996**, *8*, 1341–1390, [https://doi.org/10.1162/neco.1996.8.7.1341].
33. Tran, D.; Kucukelbir, A.; Dieng, A.B.; Rudolph, M.; Liang, D.; Blei, D.M. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787* **2016**.
34. Heo, Y.; Zavala, V.M. Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings* **2012**, *53*, 7–18.
35. Zhang, Y.; O'Neill, Z.; Dong, B.; Augenbroe, G. Comparisons of inverse modeling approaches for predicting building energy performance. *Building and E* **2015**, *86*, 177–190.
36. Kruschke, J.K. Bayesian Estimation Supersedes the t-Test. *Journal of Experimental Psychology: General* **2013**, *142*, 573–603.
37. Salvatier, J.; Wiecki, T.V.; Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2016**, *2*.
38. Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association* **1967**, *62*, 776–800.
39. Lindley, D.V. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* **1997**, *46*, 129–138.
40. Bernardo, J. Statistical inference as a decision problem: the choice of sample size. *The Journal of the Royal Statistical Society: Series D (The Statistician)* **1997**, *46*, 151–153.
41. Goldberg, M.L. Reasonable Doubts: Monitoring and Verification for Performance Contracting. ACEEE Summer Study on Energy Efficiency in Buildings; American Council for an Energy Efficient Economy, , 1996; Vol. 4, pp. 133–143.
42. Ruch, D.; Kissock, J.; Reddy, T. Prediction uncertainty of linear building energy use models with autocorrelated residuals. *Journal of Solar Energy Engineering* **1999**, *121*, 63–68.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2017 doi:10.20944/preprints201712.0032.v1

Peer-reviewed version available at *Energies* **2018**, *11*, 380; doi:10.3390/en11020380

43. Kissock, J.K.; Haberl, J.S.; Claridge, D.E. Inverse modeling toolkit: Numerical algorithms (RP-1050). *Transactions-American society of heating refrigerating and air conditioning engineers* **2003**, *109*, 425–434.
44. Reddy, T.; Claridge, D. Uncertainty of "measured" energy savings from statistical baseline models. *HVAC&R Research* **2000**, *6*, 3–20.
45. Walter, T.; Price, P.N.; Sohn, M.D. Uncertainty estimation improves energy measurement and verification procedures. *Applied Energy* **2014**, *130*, 230–236.
46. Mathieu, J.L.; Callaway, D.S.; Kiliccote, S. Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices. *Energy and Buildings* **2011**, *43*, 3322–3330.
47. VanderPlas, J. Frequentism and Bayesianism: a Python-driven primer. *arXiv preprint arXiv:1411.5018* **2014**.
48. Gelman, A.; others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis* **2006**, *1*, 515–534.
49. Hoffman, M.D.; Gelman, A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research* **2014**, *15*, 1593–1623.
50. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences. *Statistical science* **1992**, pp. 457–472.
51. Brooks, S.P.; Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* **1998**, *7*, 434–455.
52. Lange, K.L.; Little, R.J.; Taylor, J.M. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **1989**, *84*, 881–896.
53. Agnew, K.; Goldberg, M., The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures; National Renewable Energy Laboratory, 2013; chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol.
54. Booth, A.; Choudhary, R.; Spiegelhalter, D. A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. *Journal of Building Performance Simulation* **2013**, *6*, 293–318.
55. Gelman, A.; others. Analysis of variance–why it is more important than ever. *The annals of statistics* **2005**, *33*, 1–53.
56. Burkhart, M.C.; Heo, Y.; Zavala, V.M. Measurement and verification of building systems under uncertain data: A Gaussian process modeling approach. *Energy and Buildings* **2014**, *75*, 189–198.
57. Carstens, H.; Rawlins, M.; Xia, X. A user's guide to the SANAS STC WG guideline for reporting uncertainty in measurement and verification. South African Energy Efficiency Confederation Conference; , 2017.
58. Carstens, H.; Xia, X.; Yadavalli, S. Low-cost energy meter calibration method for measurement and verification. *Applied Energy* **2017**, *188*, 563–575.